Project Report – Group 6

Hang Zhang Cao, Ted Korby, Raymond Sepulveda, Dalton Bode

# Introduction:

The United States Census Bureau offers extensive data on the financial aspects of the society. This report will focus on the Annual Business Survey (ABS) Program, focusing on the Company Summary and Characteristics of Business Owners datasets of 2019. The Company Summary dataset provides data for employer businesses by a wide range of characteristics such as sector, sex, race, firm size, etc. Similarly, the Characteristics of Business Owners dataset explored the data for owners of respondent employer firms by the same range of characteristics. Based on the data given by the datasets, several questions were raised and listed below:

1. Are there any major discrepancies between races and the average annual pay?
2. What are the number of owners of mid-size firms (50-99 employees) established between 2000-2007 and 2008-2012, in MN, NY, and WI?
3. Are there any major discrepancies between certain demographics, industries, and company sizes based on the average annual payroll?
4. Is there a trend of race densities varying by geographical location? If so, where are there more or less of a certain race and what could be the explanation behind it?
5. Is there a trend of veteran status densities that vary by geographical location? If so, what are the trends?

All the data is obtained through the ABS APIs and is processed using pandas, a Python data analysis library. Visualizations are created using pandas or matplotlib (Python plotting library) to aid the discussion of answering the above questions.

# Methodology:

Question 1:

To start, create a pie chart by using the .plot.pie() method on the average number of employees data frame. The following arguments should be passed:

*subplots=True, labeldistance=None, autopct='%.2f%%', pctdistance=1.2, figsize=(8,8)*

Subplots allows for the graph to show up. The ‘labeldistance’ set to None removes the label names, but without affecting the labels for legends. ‘autopct’ labels each wedge with the percent of the area that wedge takes up and rounds it to 2 decimal places. ‘pctdisctance’ alters where the autopct labels are placed as a radius from the center of the pie chart. ‘figsize’ alters the overall size of the pie chart.

Once this is done, make sure to set the legend to the bottom right of the chart. This can be done by using the plt.legend() method, setting the parameter ‘loc’ to 3. Include a title for the chart and the result should look similar to the chart below:

The percentage of employess by race. This was gahered based on US Census data from 2018.


**Figure 1:** The percentage of employees by race. The data used to construct this visualization was gathered from the US Census Bureau from 2018.

Next, we use the avg\_pay\_by\_race data frame to create a horizontal bar chart. This can be done using the .plot.barh() method on that data frame. The following parameters were used for this method:

*y='normalized\_avg\_pay', x='Race Group', legend=None, color= ["b",'orange','g','r','purple']*

*plt.xlabel('Avg. Annual Pay rate ($1000s)')*

*plt.ylabel('Race Group')*

*plt.title('Average Annual Pay Rate ($1000s) by Race - 2018')*

The legend is set to None as it is not necessary for this chart. The colors are set to match the colors used for the pie chart.

With these parameters set, the following visualization is what you should expect:



**Figure 2:** The average annual pay rate by race in $1000s for 2018. The averages were normalized by dividing each race group’s average annual pay rate by their respective average number of employees.

Question 2:

To investigate the impact of the Great Recession on mid-size firms in New York, Wisconsin, and Minnesota, the Characteristics of Business Owners dataset is utilized along with the Company Summary dataset. By comparing the number of mid-sized firm owners between the periods of 2000-2007 and 2008-2012, we can have a grasp of the impact of the Great Recession on the financial status of mid-sized firms.

The Company Summary (ABSCS) dataset is filtered to show only the data for mid-size firms.

*pd\_cs = pd\_cs[pd\_cs['EMPSZFI\_LABEL'] == 'Firms with 50 to 99 employees']*

The Characteristics of Business Owners (ABSCBO) dataset is filtered to show data ranging from 2000-2007 and 2008-2012, stored in separate variables.

*pd\_os1 = pd\_os[pd\_os['OWNCHAR\_LABEL'] == '2000 to 2007']*

*pd\_os2 = pd\_os[pd\_os['OWNCHAR\_LABEL'] == '2008 to 2012']*

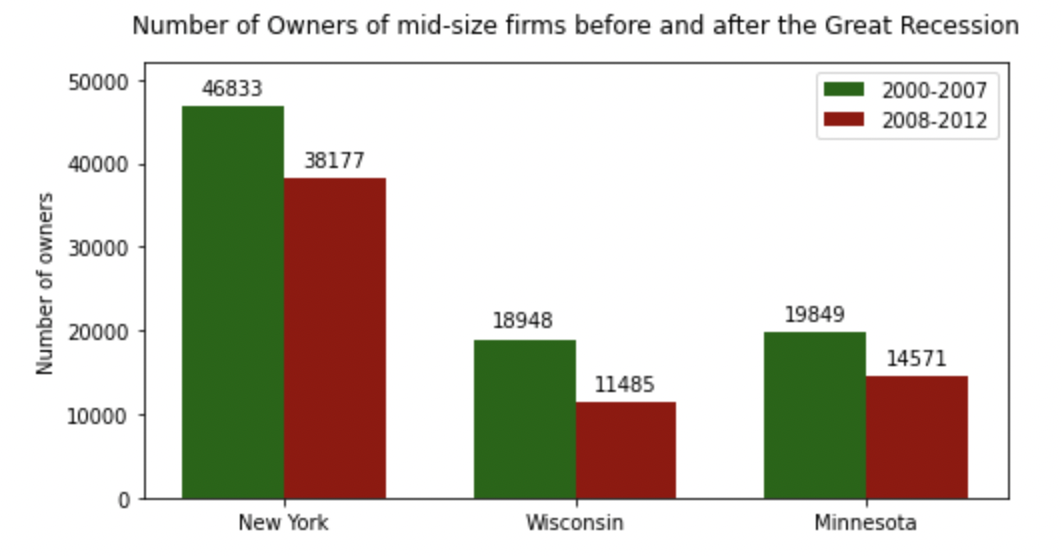
The abscbo dataset is merged with the abscs dataset on the column ‘NAME’, which includes the name of the states.

*merge = pd\_cs.merge(pd\_os1, on = 'NAME',how='inner') #2000-2007*

*merge2=pd\_cs.merge(pd\_os2, on = 'NAME',how='inner') #2008-2012*

Data for New York, Wisconsin, and Minnesota is filtered using the is in function on the ‘NAME’ column.

Clustered bar chart is then generated using matplotlib.pyplot, by plotting the state name versus the values in the ‘OWNPDEMP’ column, which represents the number of owners of respondent employer firms.



**Figure 3.** Number of owners of mid-size firms before and after the Great Recession

Question 3:

To investigate the discrepancies of the annual pay by sex, business size, and industry, the Characteristics of Business Owners dataset is used after dropping unnecessary columns. After that, the following process was used to create each graph. Although the steps are for the company size graph, the process is the same for sex and industry, with only changes in variables.

The company size data set is filtered to exclude ‘all firms’ values:

*census\_df\_dropped\_company\_size = census\_df\_dropped[(census\_df\_dropped['EMPSZFI\_LABEL'] != 'All firms')]*

The company size data set then needs to convert the annual pay column to integers so that the data can be summarized and graphed.

*census\_df\_dropped\_company\_size['PAYANN'] = pd.to\_numeric(census\_df\_dropped\_company\_size['PAYANN'])*

Next, the company size data set is grouped by the company size column and summarized by finding the average of the annual pay from each grouping. Since the averages were in tens of millions, the averages were divided by 1 million to have cleaner and easier to understand data. Then all the data from the function was stored in a variable.

*avg\_annual\_pay\_company\_size = census\_df\_dropped\_company\_size.groupby('EMPSZFI\_LABEL').mean()/1000000*

Finally, using that stored variable, the data can be plotted onto a bar chart, with the proper titles, colors, and axes applied. And the legend being removed due to it being unnecessary.

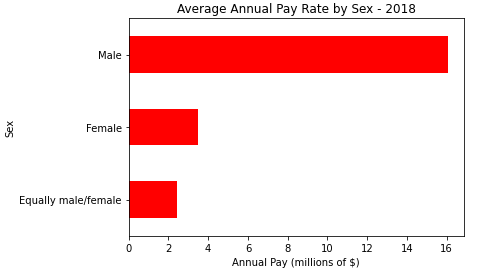
*AAPCSG = avg\_annual\_pay\_company\_size.plot(title = 'Average Annual Pay by Company Size - 2018', kind='barh', color='purple')*

*AAPCSG.set\_xlabel('Annual Pay (tens of millions of $)')*

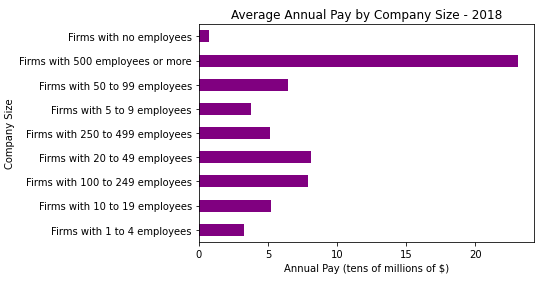
*AAPCSG.set\_ylabel('Company Size')*

*AAPCSG.get\_legend().remove()*

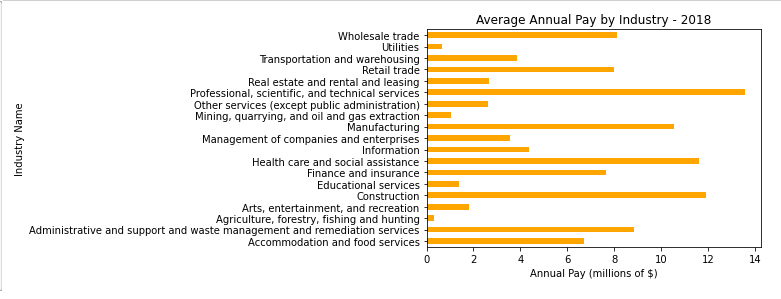
*plt.show()*



**Figure 4.** Average Annual Pay Rate by Sex in Year 2018



**Figure 5.** Average Annual Pay by Company Size in Year 2018



**Figure 6.** Average Annual Pay by Industry in Year 2018

Question 4:

***Is there a trend of race densities varying by geographical location? If so, where are there more or less of a certain race and what might be the reason behind it?***

Since we are looking for any differences in race densities by geographical location, we want to separate our data into value counts of the survey data based on the state and the race type. We do this by reducing our data to the RACE\_GROUP\_LABEL column and removing rows containing data that’s not useful for the comparison.

*tk\_races\_cut = tk\_cut.drop(columns="VET\_GROUP\_LABEL")*

*tk\_geo\_races = tk\_races\_cut[(tk\_races\_cut['RACE\_GROUP\_LABEL'] != 'Classifiable') & (tk\_races\_cut['RACE\_GROUP\_LABEL'] != 'Unclassifiable') & (tk\_races\_cut['RACE\_GROUP\_LABEL'] != 'Nonminority') & (tk\_races\_cut['RACE\_GROUP\_LABEL'] != 'Minority') & (tk\_races\_cut['RACE\_GROUP\_LABEL'] != 'Equally minority/nonminority') & (tk\_races\_cut['RACE\_GROUP\_LABEL'] != 'Total')]*

*tk\_geo\_racesgrouped = tk\_races\_geo.groupby('NAME')['RACE\_GROUP\_LABEL'].value\_counts(sort=False)*

With *tk\_geo\_racesgrouped*, we have a formatted series that contains our value counts per race for each state.

We can break this down into a more consumable format which can be later used in a detailed clustered bar graph output or pie chart.

*states = ["Minnesota", "District of Columbia", "Florida", "California", "Washington"]*

*race\_counts\_by\_state = {}*

*temp\_list = []*

*for index in tk\_geo\_racesgrouped.index:*

*if index[0] in states:*

*number = tk\_geo\_racesgrouped[index]*

*state = index[0]*

*race = index[1]*

*temp\_list.append(number)*

*if len(temp\_list) == 5:*

*race\_counts\_by\_state[state] = temp\_list*

*temp\_list = []*

The previous block of code yields us a dictionary list filled with the different states and their individual race counts without a direct correlation, but the order of each race is the same due to the sorting of the value\_counts function being set to False.

With this, we can construct a DataFrame that takes in the values of the counts, indexed by the states with each column being set as their respective order as in the original value counts order. We transpose this data to make the races act as our categories and states as our comparative measure.

*races\_df = pd.DataFrame(data=race\_counts\_by\_state.values(), index=race\_counts\_by\_state.keys(), columns=['American Indian and Alaska Native', 'Asian', 'Black or African American', 'Native Hawaiian and Other Pacific Islander', 'White'])*

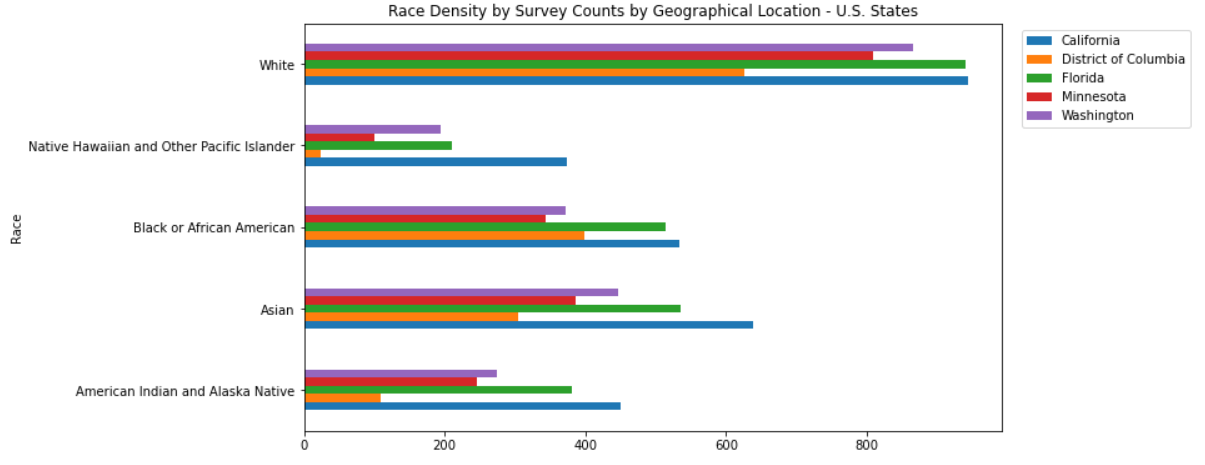
*races\_df = races\_df.T*

From this new DataFrame, we can plot our data to make our analysis.

*races\_plot = races\_df.plot(kind="barh",figsize=(10, 6), legend=1)*

*races\_plot.legend(loc=9,bbox\_to\_anchor=(1.15, 1.0))*

*races\_plot*



**Figure 7.** Race Density by Survey Counts by Geographical Location – U.S. States

*Question 5:*

***Is there a trend of veteran status densities that vary by geographical location? If so, what are the trends?***

Since we are looking for differences in veteran status densities based on geographical location, we can follow the same basic steps from the previous question on race densities based on geographical location. The main differences are the column and row data that we will be including, for this example we will not look at ‘Classifiable’ and ‘Unclassifiable’ records due to the ambiguous meaning.

*tk\_vet\_specific = tk\_cut[(tk\_cut['VET\_GROUP\_LABEL'] != 'Classifiable') & (tk\_cut['VET\_GROUP\_LABEL'] != 'Unclassifiable')]*

*tk\_vet\_geo = tk\_vet\_specific.drop("RACE\_GROUP\_LABEL", axis=1)*

*tk\_geo\_vetgrouped = tk\_vet\_geo.groupby("NAME")["VET\_GROUP\_LABEL"].value\_counts(sort=False)*

I chose to keep the total counts in the veteran status comparison due to the removed rows having a large impact on the amount of data present, so we may find it helpful to know how much of the data at hand truly represents the population surveyed.

Following the same steps as in the race densities question, we create dictionaries to be put into a DataFrame which can be used to create visuals that will help us in answering our question.

*states = ["Minnesota", "District of Columbia", "Florida", "California", "Washington"]*

*vet\_counts\_by\_state = {}*

*temp\_list = []*

*for index in tk\_geo\_vetgrouped.index:*

*state = index[0]*

*if state in states:*

*number = tk\_geo\_vetgrouped[index]*

*temp\_list.append(number)*

*if len(temp\_list) == 4:*

*vet\_counts\_by\_state[state] = temp\_list*

*temp\_list = []*

*vets\_df = pd.DataFrame(data=vet\_counts\_by\_state.values(), index=vet\_counts\_by\_state.keys(), columns=['Equally veteran/nonveteran','Nonveteran','Total','Veteran'])*

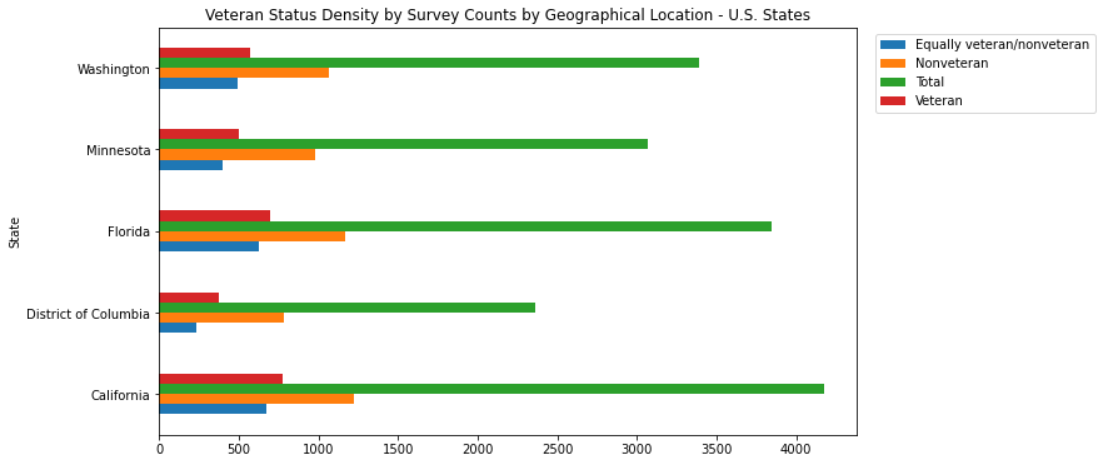
*vets\_df.T*

With this, we have the data necessary to create the visuals necessary to help answer the questions at hand.

*vets\_plot = vets\_df.plot(kind="barh",figsize=(10, 6), legend=1)*

*vets\_plot.legend(bbox\_to\_anchor=(1.35, 1.0))*

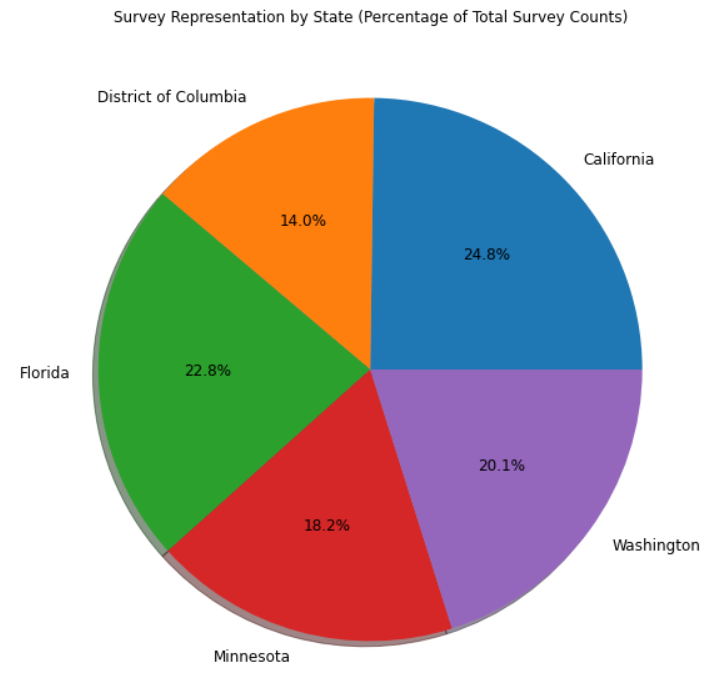
*vets\_plot*



**Figure 8.** Veteran Status Density by Survey Counts by Geographical Location – U.S. States

It

*vets\_pie = vets\_df.plot.pie(y="Total", ylabel='', title= "Survey Representation By State (Percentage of Total Survey Counts)", figsize=(10,10), legend=False, autopct='%1.1f%%', shadow=True, fontsize=12)*



**Figure 9.** Survey Representation by State (Percentage of Total Survey Counts)

# Discussion:

Answering those questions or indicate if more research may be needed. If more research is needed, suggest what research may need to be done.

Figures 1 and 2 show major discrepancies between the percentage of employees by race group in the US. In Figure 1, with 86% of employees in the US for the year 2018 being White, and the next closest race group, Asian, comprising another 10% of employees, one would assume with an overwhelmingly higher employee population, this should impact the annual pay rates.

This was the case for the smallest employee populations, which were the Native American and Hawaiian/Pacific Islander race groups, whose small populations increased their average pay rate per employee to be comparable to other race groups. This was not the case for the Black race group, whose average annual pay rate was significantly smaller than every other race group, even after these numbers were normalized by their relatively small employee population size. What is equally questionable is the White race group’s average annual pay rate being the largest average pay rate overall, even after normalization, as the pay rate should have followed a similar trend to the Asian race group’s average annual pay rate.

This is a clear indication of discrimination, in which the injured parties should consider renegotiating their contracts for higher pay, or file lawsuits to offset these discrepancies. Business owners should restructure their hiring processes to take these demographics into consideration, as well as normalizing pay rates during contract negotiations. If these issues arise from lack of experience or education, employers could offer resources for individuals rejected from the hiring process on where to find information related to what their company expects their employees to know as a foundation for that position.

It is evident from Figure 3 that the number of mid-size firm owners declined after the Great Recession. This data can be used as an indicator to evaluate the financial situation during that period. However, the dataset did not explicitly define the meaning of “owner.” This leaves ambiguity as owner can either be an individual holding 100% share of the stock or someone in the board of directors. More direct data such as the number of firms can be introduced to provide further insight into this question.

It has been made evident in Figure 4 that there is a major discrepancy for the sexes when it comes to average annual pay. When it comes to equally male/female or female businesses, they have an average annual pay of 2 and 4 million dollars, respectively. On the other hand, male owned businesses are paid 4 times more annually compared to other sexes, with an average annual pay of 16 million dollars. Further investigation into other aspects of the businesses and the business owners (education, sales, etc.) may provide insight into why the discrepancy is so large between the sexes.

Based on the findings from Figure 5, there is also one major discrepancy of average annual pay when it comes to business size. Businesses that have over 500 employees have an average annual pay of over 20 million dollars. Whereas businesses with less than 500 employees have an average annual pay of less than 10 million dollars. Interestingly enough, the discrepancies of companies that have less than 500 employees are not as significant, with discrepancies only ranging from half a million to about 7 million dollars at most. Further investigation of the smaller companies specifically may provide some insight into why the discrepancies are not as significant.

Based on the findings from Figure 6, discrepancies can vary depending on the industry being investigated. From the census, top industries had an average annual pay of over 10 million dollars. On the other hand, other industries had an average annual pay of 2 million dollars or less. Further investigation, especially into industries that had an average annual pay of 2 million dollars or less, may be necessary to see why the pay is lower compared to other industries.

There are a few takeaways from Figure 7 visual, the most obvious of which is that across all states the most densely populated race is White. This is most inherently due to the United States of America being constructed and led by several white demographics from the western end of the Eastern Continents. Relating this to overall population counts or survey counts by all races, each state tends to have a rather consistent racial profile regardless of region but usually has one race as more common in comparison to other regions based on historical association such as cultural background. Utilizing this information, we can make an educated assumption on the fact of white majority in nearly all states/regions and perhaps historical/cultural affiliation to a region has an impact on the racial diversity of a region.

There is an interesting bit of information that pops out about the Native Hawaiian and Other Native Pacific Islanders in Figure 7. There is a considerably larger population density relative to the west coast than there is to the eastern coast states. Although, there is a slight outlier within that assumption such that Florida has the second largest population density of that race. This could mean that the temperate of a region may have an impact on the racial diversity of a region.

Using the assumptions from the previous two discussion points, this data could be used to increase awareness of certain racial related debates and may even heighten research and educational tools on topics related to racial-geological relations. These observations may also be used if targeting an audience may be useful, such as presidential campaigns operating in swing states.

Based on the findings of Figure 8, we noticed that all states/regions have more non-Veteran surveys than Veteran or Equally Veteran/Non-Veteran surveys. This could be due to the nature of the nation being a free nation, allowing the population to choose what they want to do with their lives, and many choose not to join a branch of the US military.

Looking further into Figure 8 and using Figure 9 as a reference to the survey count differences, we can see that the District of Columbia has roughly over half of the veteran status density by survey counts of California, but by relative non-Veteran counts it has more. This reveals an interesting trait about the region such that veterans tend to not reside in the region of District of Columbia. When combining Veteran and Equally Veteran/non-Veteran survey counts for each state in Figure 8, most states fall equal to or greater than the non-Veteran survey counts except for District of Columbia which is short of equal to meaning there are on average less veterans in that region, assisting in the analysis.

Conclusion:

From information on race/gender demographics to average annual pay rates for any given industry, these are a few examples in which businesses can derive insight from the US Bureau Census data to improve their internal processes or inform future job seekers on choosing a career path that simultaneously suits their passion and financial needs.

The data at hand is not complete as it doesn’t accurately reflect the total populations of any set of demographics but does give a strong estimate based on a subset of the populous through a randomized survey system. We would recommend that the usage of data from multiple data sources following the same data structure would allow for a truer representation of the population and its varying demographics.

# Works Cited:

API reference. API Reference - Matplotlib 3.5.1 documentation. (n.d.). Retrieved January 15, 2022, from <https://matplotlib.org/stable/api/index>

Hoefler, P. (n.d.). *Pandas.dataframe*. pandas.DataFrame - pandas 1.3.5 documentation. Retrieved January 15, 2022, from <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html>

US Census Bureau. (2021, October 14). *Annual Business Survey (ABS) APIs*. Census.gov. Retrieved January 14, 2022, from <https://www.census.gov/data/developers/data-sets/abs.html>